Modelling binaural detection of speech stimuli in complex reverberant environments

Tobias Weller and Jörg M. Buchholz
Department of Linguistics, Macquarie University, 2109 Macquarie University NSW, Australia
Virginia Best
National Acoustic Laboratories, 2109 NSW, Australia

Summary
In listening experiments that assess spatial hearing (e.g. localization) in the presence of background noise it is important to ensure audibility of the target stimulus. Target audibility could be easily controlled using knowledge about the masked thresholds in the respective paradigm. However, if a large number of acoustic conditions as well as target sounds are considered, individual psychoacoustic threshold measurements are not feasible and alternative methods need to be applied. In this study a binaural auditory model is applied to predict masked thresholds for wide-band non-stationary stimuli. This model closely follows the approach of Hant and Alwan (Speech Communication, 2003, Vol. 40, pp. 291-313) [7]. In the first stage of the model an auditory front-end generates an internal representation of the stimuli in both ears. The auditory front-end includes head-related transfer functions, auditory bandpass filtering, squaring, temporal integration, logarithmic compression and additive internal noise. In the second stage, a decision device combines $d'$ information across time, frequency and ears and provides an estimate of the masked threshold. Three different methods of combining information across ears were compared: A better-ear approach, a cross-ear glimpsing approach and an approach using binaural integration. The model was verified using psychoacoustic masked threshold data of a detection paradigm that considered a target word presented from 15 different locations in a reverberant multi-talker background. The model predictions were in very good agreement with the measured masked thresholds. It is concluded that this model is suitable to control audibility of a target stimulus in a complex reverberant environment. Moreover, the model was used to systematically analyze the effect of head-shadow, signal spectrum, auditory sensitivity and room reverberation on target audibility.

PACS no. 43.66.Ba, 43.66.Dc, 43.66.Pn

1. Introduction
It is widely acknowledged that hearing aids typically have a detrimental effect on cues that the human auditory system uses for spatial hearing (localization, scene analysis etc.) [13]. To allow conclusions about the everyday performance of hearing aids, experiments studying this effect should ideally be conducted under reproducible, realistic conditions. Virtual acoustic environments (VAEs) such as loudspeaker-based room auralization [4] offer this kind of conditions [9].

Conducting an experiment that assesses spatial hearing in a VAE requires thorough control of the audibility of the target stimulus. If the target is played below or close to the masked threshold, audibility will be measured instead of spatial hearing ability. In a complex background, masked thresholds will not only depend on the background noise but also the target’s location with respect to the listener. It is therefore necessary to ensure target audibility for all target locations individually.

Measuring target audibility for multiple target locations can be a tedious and time-consuming task. A detection model predicting masked thresholds would have the potential to greatly reduce the time subjects have to spend on a spatial assessment task. Considering realistic everyday acoustic scenarios, such a model has to be able to cope with complex stimuli such as speech targets in a multitalker background.

Temporal integration in auditory detection as well as loudness perception is typically modelled by applying long temporal integration constants of about 80–300 ms, while auditory temporal resolution, such as measured in gap-detection experiments, is usually
modelled using very short temporal integration constants of about 3–5 ms [10]. To solve this paradox, Viemeister and Wakefield [14] proposed a multi-look approach to modelling detection, which integrates information over multiple "looks" of relatively short duration (≈ 5 ms). Similarly, Florentine and Buus [5] proposed a multiband excitation model to predict thresholds of wideband stimuli. Hant and Alwan [7] then combined these two approaches to apply the multiple-look concept to predict thresholds for wideband non-stationary signals of varying duration. Their model was successful in predicting thresholds for plosive bursts in speech shaped noise as well as masked discrimination thresholds for CV syllables.

In the present study an extension to the model proposed by Hant and Alwan is presented that allows the application of binaural stimuli. This extended model is used to predict data from a study measuring masked thresholds of a speech target in a reverberant multitalker environment [15]. Fifteen target locations are considered with varying direction as well as target-receiver distance to study the effect of direction and amount of reverberation on auditory detection. Three different approaches to the binaural extension are used and their predictions are compared to the experimental data. Furthermore, the model is used to analyze the acoustic and auditory cues that are mainly affecting the observed detection task.

2. Model structure

The model is divided into two stages, the auditory front end and the decision device. The auditory front end is a signal processor that generates an internal representation of the stimuli. Based on this internal representation the decision device computes and combines information across time, frequency and both ears. The masked threshold is defined as the SNR that yields an overall matching a theoretical value corresponding to the task applied in the psychoacoustic experiment.

2.1. Auditory front end

The processing steps of the auditory front end are shown in Fig. 1. First the stimuli arriving at the listener’s left and right ear are processed separately by an auditory filterbank. The resulting narrowband signals are half-wave rectified by squaring. After that, the signals are integrated over time, logarithmically compressed. Finally, Gaussian noise is added to each channel, representing the internal noise of the auditory system.

Center frequencies range from 105 Hz to 7411 Hz.

The narrowband signals are processed by a temporal integrator using a flat window with raised-cosine skirts. The flat part of the window is 4 ms long and the skirts have a duration of 1 ms each. The window is applied every 5 ms resulting in an overlap of 1 ms between temporally adjacent windows. Due to the use of raised-cosine skirts and the overlap of 1 ms the stimulus is weighted equally over its duration.

After temporal integration, the signal is logarithmically compressed before internal noise is added. The internal noise is modelled as Gaussian noise with zero mean and frequency-dependent variance \[ \sigma_{\text{int}}^2(k) \], where \( k \) is the ERB number. Based on Hant and Alwan [7] the values of the variance were given by the sigmoidal fit

\[
\sigma_{\text{int}}^2(k) = 16.62+7.88 \left( -\frac{1}{2} \cdot \frac{1}{2} \cdot 1 - e^{k-21.8} \right) \]

(2)

2.2. Decision device

In the second stage of the model a decision device provides an estimate of the detectability \( d' \) of the target based on the time-frequency representation at the output of the auditory front end. First, a \( d' \) value is calculated for each time-frequency look independently. The overall \( d' \) is then determined by an optimal combination of individual \( d' \) values across time, frequency and both ears. The masked threshold is found by iteratively adjusting the SNR between target and masker signals until the combined \( d' \) resulting from the decision device matches the theoretical \( d' \) of the given psychoacoustic task.

The decision device compares the internal response to the masker alone (denoted M) with the internal response to a mix of the masker and the target (denoted S). Note that M and S refer to the internal representation of the stimuli, i.e. the output of the auditory front end at either the left or right ear. The
detectability of the target in a single time-frequency look is then given by

$$d'_{nk} = \frac{|\mu_{S_{nk}} - \mu_{M_{nk}}|}{\sqrt{\sigma_{S_{nk}}^2 + \sigma_{M_{nk}}^2}} \bigg/ 2$$

(3)

where $n$ and $k$ are the time index and the frequency index and $\mu$ and $\sigma^2$ denote the corresponding mean and variance respectively. The mean and variance values for each frequency look are determined from 2000 samples of the internal responses $S$ and $M$. If the masker stimulus has a much longer duration than the target stimulus, $S$ can be obtained by adding the target stimulus to a different, randomly chosen part of the masker for each sample.

The optimal combination of detectability values across time and frequency is then given by

$$d' = \sum_{n=1}^{N_t} \sum_{k=1}^{N_f} (d'_{nk})^2$$

(4)

with $N_t$ and $N_f$ being the number of time looks and frequency looks, respectively.

To account for decreased detector efficiency when dealing with wideband and/or long-duration stimuli Hant and Alwan [7] introduced a weighting function $w$. This binary weighting function multiplies time-frequency looks where the difference $|\mu_{S_{nk}} - \mu_{M_{nk}}|$ is below a threshold $\theta$ with a factor of 0. This extends Eq. (4) to

$$d' = \sum_{n=1}^{N_t} \sum_{k=1}^{N_f} w_{nk} (d'_{nk})^2$$

(5)

with

$$w_{nk} = \begin{cases} 1 & \text{if } |\mu_{S_{nk}} - \mu_{M_{nk}}| \geq \theta, \\ 0 & \text{if } |\mu_{S_{nk}} - \mu_{M_{nk}}| < \theta. \end{cases}$$

Similar to the variance of the internal noise, the threshold $\theta$ varies with center frequency. The threshold values suggested by Hant and Alwan [7] are given by

$$\theta(k) = 3.81 + 2.39 \left( \frac{1}{2} + \frac{1}{2} \cdot \frac{1 - e^{-0.1615k}}{1 + e^{-0.1615k}} \right).$$

(7)

Equation (5) provides a method to determine the detectability of a target stimulus in the presence of a masker. However, Eq. (5) can only be used if the stimuli are monaural. We compared three different methods of extending the model to predict the detectability of binaural stimuli:

(a) Detection using the better ear
(b) Detection using cross-ear glimpsing
(c) Detection using binaural combination

Each of these methods uses a different way of integrating information across ears. Refer to Fig. 2 for an overview over the three different methods.

2.2.1. Detection using the better ear

In this approach $d'$ is first determined for each ear individually following Eq. (5). The actual overall detectability is then defined to be the maximum of the two individual values:

$$d' = \max \{d'_{L}, d'_{R}\}$$

(8)

where indices $L$ and $R$ refer to the left and right ear, respectively.

2.2.2. Detection using cross-ear glimpsing

In the cross-ear glimpsing approach, each sample of the internal representations $S$ and $M$ that are passed on to the decision device is a mixture of time-frequency looks from the right and left ear. At the
output of the auditory front end the SNR is estimated at each ear and for each time-frequency look using
\[
\hat{\text{SNR}}_{L/R}^{nk} = S_{L/R}^{nk} - M_{L/R}^{nk}.
\] (9)
Each time-frequency look is then chosen from the ear with the better SNR at time index \(n\) and frequency index \(k\) so that the internal representations are composed of time-frequency looks from both ears for different instances in time and frequency:
\[
\begin{align*}
S_{nk} &= S_{nk}^L \quad \text{if } \hat{\text{SNR}}_{L}^{nk} \geq \hat{\text{SNR}}_{R}^{nk} \\
M_{nk} &= M_{nk}^L \\
S_{nk} &= S_{nk}^R \quad \text{if } \hat{\text{SNR}}_{L}^{nk} < \hat{\text{SNR}}_{R}^{nk} \\
M_{nk} &= M_{nk}^R
\end{align*}
\] (10)-(11)
\(S_{nk}\) and \(M_{nk}\) are then passed on to the decision device and evaluated like a monaural internal representation following Eqs. (3) and (5).

2.2.3. Detection using binaural combination
In this approach, the individual detectability values \(d'_{nk}\) for each time-frequency look are first determined and multiplied by the weighting function for each ear separately. The remaining information after application of the weighting function is then added over both ears, resulting in an overall detectability
\[
d' = \frac{1}{N_t} \sum_{n=1}^{N_t} \sum_{k=1}^{N_f} \sum_{i=\{L,R\}} w_{nk,i} \left( d'_{nk,i} \right)^2.
\] (12)
The index \(i\) is here used to refer to either the left ear or the right ear, respectively.

Note that the general sensitivity of the model with all three binaural extensions is defined by the parameter fit of Hant and Alwan as given in Eqs. (2) and (7).

3. Model predictions of masked thresholds for speech in speech stimuli in a reverberant environment
3.1. Stimuli and experimental procedure
The model was used to predict data of an experiment measuring masked thresholds for a speech target in a reverberant multi-talker environment [15]. In this experiment a cafeteria-like room was simulated using room simulation software (ODEON, [12]), an auralization toolbox for MATLAB (LoRA, [4]) and a 41-channel loudspeaker array located in an anechoic chamber.

A female voice saying the word "two" was used as the target and the masker was composed of seven concurrent conversations between two persons each.

Figure 3. Top-down view of the virtual room used in the experiment. The transparent head represents the listener, green heads indicate the target positions in 1, 2 and 4m distance and red heads indicate the positions of sources contributing to the masker.

Figure 4. The input stimuli for the model are generated by first filtering the 41 loudspeaker signals used in the behavioral experiments with HRTFs. The HRTFs were measured from each loudspeaker to each ear. The filtered signals are then summed up for both ears individually, resulting in the input stimulus.

Masked thresholds were measured for 15 different locations of the target within the virtual room (see Fig. 3), while the positions of the listener and the distractors remained unchanged. Masked thresholds were measured using an adaptive 1-up 2-down procedure defining the threshold as the 70.7% correct point on the psychometric function [8]. Furthermore, the experiment employed a three-alternative forced-choice (3-AFC) task setting the detectability goal for the model to \(d' = 1\) [6]. Nine normal-hearing subjects with an average age of 29.3 years participated in the experiment.

The input stimuli for the model were generated such that they would resemble the signals at the eardrum of the subjects in the experiment as closely as possible. The 41 loudspeaker signals used in the experiment were filtered with head-related transfer functions (HRTFs) measured from each loudspeaker to the two ears of a Head And Torso Simulator (HATS). The filtered signals were then summed up for each ear separately to obtain appropriate input stimuli for the model. A diagram of this process is depicted in Fig. 4.
3.2. Experimental results and model predictions

The experimental results presented in [15] and the corresponding model predictions are shown in Fig. 5. Masked thresholds are plotted as a function of the direction of the target with respect to the listeners look direction. The top, middle and bottom panels show results for targets in 1 m, 2 m and 4 m distance, respectively. For most of the target locations, model predictions are within ±σ of the measured thresholds.

The better-ear (section 2.2.1) and the binaural combination model (section 2.2.3) approaches overestimate thresholds for targets at the front (0°) in 1 and 2 m distance of the listener. The cross-ear glimpsing approach (section 2.2.2) performs better at these targets locations, however, it tends to underestimate thresholds for lateral target locations (all distances) and for the target at 180° and 2 m distance.

In Table I the overall performance of the three model approaches is shown in terms of the root mean square (RMS) error between the predictions and the experimental results. Over all 15 target positions the binaural combination approach performs best and the cross-ear glimpsing approach performs the worst. However, performance highly depends on target distance: for a target distance of 1 m, the cross-ear glimpsing approach produces the lowest RMS error, while for a target distance of 4 m it produces a much larger RMS error than the other two approaches.

4. Discussion

Figure 5 shows that masked thresholds heavily depend on the target direction when the target is at a distance of up to 2 m. This is well reflected in the model predictions, as both the masked thresholds and the model predictions vary with target direction by about 8 dB SNR at those distances. This influence of target direction on the masked thresholds is considerably lower when the target is at a distance of 4 m. Here the results only vary by about 2.5 dB SNR. Additionally, increasing the distance between source and listener also increases the masked thresholds overall.

The direction dependency for targets close to the listener seems to be mainly explained by the directivity of the head-related transfer functions (HRTFs). The HRTFs integrated over both ears and in two example frequency regions are plotted in Fig. 6 as functions of azimuth angle. While the HRTFs vary only

<table>
<thead>
<tr>
<th>target distance</th>
<th>1 m</th>
<th>2 m</th>
<th>4 m</th>
<th>all pos.</th>
</tr>
</thead>
<tbody>
<tr>
<td>better-ear</td>
<td>2.37</td>
<td>1.45</td>
<td>1.04</td>
<td>1.69</td>
</tr>
<tr>
<td>cross-ear glimps.</td>
<td>1.34</td>
<td>1.98</td>
<td>2.47</td>
<td>1.95</td>
</tr>
<tr>
<td>binaural comb.</td>
<td>1.76</td>
<td>1.28</td>
<td>1.07</td>
<td>1.39</td>
</tr>
</tbody>
</table>

Table I. RMS error between the model predictions and the experimental data. All values are in [dB SNR].

Figure 5. Masked thresholds measured in the experiment and model predictions for different approaches to a binaural decision device plotted over target direction. Masked thresholds are averaged over all participating subjects and the error bars indicate ±1 standard deviation. The top, middle and bottom picture show results for targets in 1 m, 2 m and 4 m distance, respectively.

Figure 6. HRTFs integrated over both ears between 1.6 kHz and 2.1 kHz (black squares) as well as 5.6 kHz and 7.8 kHz (gray circles) plotted over azimuth angle. The results are normalized with respect to their 0° point.
Figure 7. Direct-to-reverberant ratio averaged over all target locations at one specific distance plotted over target distance. Linear interpolation locates the critical distance at $d_c = 2.75$ m.

by 4 dB in the frequency region between 1.6 kHz and 2.1 kHz (corresponding to frequency channels 18 and 19 in the filterbank of the auditory front end of the model) the HRTFs between 5.6 kHz and 7.8 kHz (corresponding to frequency channels 29 – 31) vary in a range of 10 dB. The maxima of the HRTFs (between $\pm 45^\circ$ and $\pm 112.5^\circ$) correlate with the minima of the masked threshold functions for 1 m and 2 m target distance. This suggests that high frequency components play an important role in the detection of the close targets. This is also supported by the auditory system-internal sensitivity plots given in Fig. 8 as further discussed below.

Going from 2 m to 4 m distance from the listener highly decreases the influence of target direction and overall increases masked thresholds. Moving further away from a sound source in a room with reflective walls usually increases the amount of reverberant energy compared to the energy of the direct sound. Note that all target stimuli were normalized with respect to their RMS sound pressure level, i.e. the overall stimulus energy was not altered by increasing the distance between source and listener. Figure 7 depicts the average direct-to-reverberant (DR) ratio of sources in the simulated room plotted over the distance between the sources and the listener. It shows that the DR ratio drops from 9.3 dB at 1 m to $-3.68$ dB at 4 m. The critical distance, where direct energy equals reverberant energy, is at $d_c = 2.75$ m. Hence, whereas in the 1 m and 2 m target conditions the direct sound energy is mainly determining audibility, in the 4 m target conditions the reverberant energy is dominant and the directivity of the HRTF of the target signal is less important.

Figure 8 provides further analysis of the effect of target distance. It shows different stages of the decision device of the model (in this case the cross-ear glimpsing approach), from top to bottom: a) the target stimulus waveform, b) individual detectability values $d'_{nk}$ for all looks $n$ and $k$, c) weights $w_{nk}$ and d) the product $d'_{nk} \cdot w_{nk}$. The left column contains the figures corresponding to the target at $-90^\circ$, 1 m and the right column contains the figures corresponding to the target at $-90^\circ$, 4 m. For both targets, the SNR was set to the respective predicted threshold such that the overall detectability was $d' = 1.26$.

Figure 8 (left column) illustrates that for a close target almost the entire energy contributing to its detectability is contained in the "echo-free" onset of the stimulus, i.e. the voiceless consonant "t" of the word "two". Most of the time-frequency looks that remain after applying the weighting function $w_{nk}$ are located in frequency bands 29 to 31 (5.6 – 7.8 kHz). Hence, the frequency region where the HRTF is highly directional (see Fig. 6) mainly determines the detectability of the target, explaining the directional dependency of the masked threshold at short distances. For the target well beyond the critical distance, detectability is more strongly affected by reverberant energy in the harmonic part of the stimulus, located in frequency bands 18 and 19 (1.6 – 2.1 kHz) and to lesser extent in frequency bands 6 and 7 (0.3 – 0.43 kHz). As was shown in Fig. 6, the HRTFs vary much less over direction for lower frequencies, suggesting that not only the more diffuse characteristic of the far target signals account for the diminishing effect of target direction, but also the shift of the main auditory sensitivity to lower frequencies.

The model predicts both higher overall thresholds and decreased effect of direction with increased target distance with all three proposed decision devices. However, the decision devices differ in sensitivity. The better-ear approach offers the lowest sensitivity because in any case, time-frequency looks from one ear only are considered. The cross-ear glimpsing approach also only considers looks from one ear, however for each look the ear with the best SNR is chosen, while in the better-ear approach time-frequency looks from the overall better ear will contribute less to the detectability than looks from the weaker ear. Even though the binaural combination approach considers looks from both ears for detectability, it ranks between the two other approaches in sensitivity. This is due to the fact that in this approach, individual detectability values are weighted separately for each ear, thus only time-frequency looks with high SNR will be taken into account.

In terms of the RMS error between the model predictions and the experimental data the binaural combination approach offers the best fit. However, despite the good agreement of the model predictions with the experimental data, no statement about the physiological plausibility of the chosen model approaches can be made due to their highly functional character. Future extensions to the model may consider more auditory system-motivated approaches based on interaural cross-correlation analysis [2] or equalization-cancellation theory [3, 1].
Figure 8. Illustration of different stages in the decision device (cross-ear glimpsing) for two different targets: A target at $90^\circ$, 1 m (left column) and a target at $90^\circ$, 4 m (right column). From top to bottom: a) the target stimulus waveform, b) individual detectability values $d'_{nk}$ for all looks $n$ and $k$, c) weights $w_{nk}$ and d) the product $d'_{nk} \cdot w_{nk}$.

5. Summary and Conclusions

The auditory detection model proposed by Hant and Alwan was successfully extended to allow the application of binaural stimuli. Three different decision devices were implemented and their predictions compared to experimental data: a better-ear approach, a cross-ear glimpsing approach and a binaural combination approach.

An analysis of the model and the stimuli suggests that the masked thresholds are mainly influenced by the directivity introduced by head-shadow (as observed in HRTFs) and the direct-to-reverberant (DR) ratio of the target. With all three decision devices the model correctly predicts these effects of target location on the masked thresholds.

Overall the model produces predictions with the lowest RMS error when the binaural combination decision device is applied. The RMS error is small enough to allow the application of the model for the purpose of control of target audibility. For the testing of spatial hearing with (aided) hearing impaired listeners the model needs to be extended to include the effect of hearing loss.

However, although further confirmation should be provided by comparing measured and modelled detection data in additional acoustic environments, the fact that the present data was successfully predicted using the original model parameters derived by Hant and Alwan [7] is highly promising.

Acknowledgement

The authors acknowledge the financial support of the HEARing CRC, established and supported under the Cooperative Research Centres Program - an initiative of the Australian Government.

References


